THE POWER-SERIES ALGORITHM FOR POLLING SYSTEMS WITH TIME LIMITS

J. P. C. Blanc
Tilburg University, CentER for Economic Research,
P.O. Box 90153, 5000 LE Tilburg, The Netherlands.

February 22, 2008

Abstract

This paper deals with evaluation and optimization of polling systems with time limits. Performance measures are evaluated with the power-series algorithm, a flexible technique for computing performance measures for multi-queue systems. The constant time limits are approximated by Erlang distributed variables. The algorithm is extended to compute derivatives of performance measures. This allows for optimization of cost functions with respect to the mean values of the time limits by gradient methods. Several properties of the optimal time limits are revealed by the numerical solution of various optimization problems.

1 Introduction

In many communication, production and other systems several types of jobs compete for access to a single service facility, e.g., a communication channel or a machine. Such systems are often modeled as polling systems. These are multi-queue systems with a single server, who attends to the jobs in the various stations according to some visit rule and some service rules. The visit rule determines the order in which the stations are visited by the server. The service rules determine the number of services that the server is allowed to perform during the subsequent visits to the stations. The choice of these rules may partly be limited by physical constraints, but can otherwise be used to control the quality of service provided to each of the job types.
Polling systems are generally hard to analyse. For some systems, e.g. with exhaustive or gated service, it is possible to derive sets of linear equations that determine the moments of the waiting time distributions, cf., e.g., [1], [11]. Some two-queue models can be solved analytically for a larger class of service disciplines, but to obtain numerical data from these solutions may require substantial effort, cf. [8] which deals with a two-queue system with exponentially distributed timers. In most other cases, performance measures can only be approximated by numerical techniques based on the solution of balance equations for state probabilities, or estimated by simulation. The power-series algorithm (PSA) is one of the available methods. It requires a Markov representation of the queueing process, possibly with the aid of some supplementary variables. It is based on power-series expansions of the state probabilities in terms of a parameter of a system for (recursively) solving the global balance equations satisfied by these probabilities. It is a flexible method which is applicable to a wide class of multi-queue/multi-server models, with Markovian Arrival Processes and phase-type service time distributions. The PSA is also suitable for optimization purposes, since it allows the computation of derivatives of performance measures with respect to system parameters and control variables. For moderately sized systems, the PSA favorably compares with simulation and numerical methods based on truncation of the state space. This is mainly so because the PSA involves recursive schemes and allows the application of the so-called ϵ-algorithm which strongly improves the convergence of the power series, cf. [2]. Since the memory requirements rapidly grow with the number of queues, the PSA can only produce accurate results for systems with a limited number of queues. The main contribution of the PSA lies in studying the interaction between queues on a reduced scale and in developing and testing approximations of performance measures and optimal val-

ues of control variables for systems of a larger size. [3] reviews the PSA in its generality. The applicability and complexity of the PSA for polling systems with various visit and service rules have been discussed in [2]. The computation of derivatives with the aid of the PSA has been described in a general context in [6]. The latter paper considers as an example the problem of optimizing cost functions for polling systems with respect to the parameters of the service rules, in that case so-called Bernoulli schedules. In [7], the PSA has been used to test the quality of several approximative approaches for determining the optimal job limits in cyclic polling systems.

An alternative technique for computing performance measures of polling systems is the discrete Fourrier transform method. It is based on relations for the generating function of the queue length distribution at various imbedded time instants. [10] applies the discrete Fourrier transform method to cyclic polling systems with non-preemptive, time-limited service. This method allows for general service and switching time distributions. However, the time limits have to be approximated by exponentially distributed timers. In the present paper, we develop the PSA for this type of polling models. Although the PSA can deal with phase type distributions for the service and switching times and with Markovian arrival processes, cf. [13], the recursions of the PSA will be presented for exponential distributions. In this way, the structure of the recursions w.r.t. the time limits will appear more accentuated. The results of [10] will be extended in several directions. Firstly, we will incorporate Erlang distributed timers in order to approximate the constant time limits more closely. Secondly, our model is readily extended to general periodic visit orders, cf. [4]. Finally, we will consider the problem of choosing the values of the time limits such as to minimize some cost function, with or without restrictions on the time limits. For this purpose, the computation scheme of the PSA includes recursions for the coefficients of derivatives of performance measures with respect to the transition rates of the time limits. The main differences with earlier implementations of the PSA are the additional states required to approximate the service discipline and the optimization w.r.t. unbounded decision variables not explicitly contained in the recursions, which, however, are better amenable to constraint optimization than e.g. Bernoulli schedules. There are no practically useful error bounds for computations with the PSA together with the ϵ-algorithm — the same holds for the discrete Fourrier transform method and truncation methods — but inspection of results produced with an increasing number of terms of the power series gives a good indication of the attained accuracy, cf. [2].

The paper is organized as follows. In Section 2 the polling model will be described in more detail, and the necessary notations will be introduced. Section 3 contains the global balance equations for the queue-length process extended with several supplementary variables. The recurrence relations of the computation scheme of the PSA for the state probabilities are given in Section 4. Here, the influence of the number of phases of the Erlang distributions for the timers is illustrated. Section 5 extends the recursive scheme of the PSA to the computation of derivatives w.r.t. the transitions rates of the timers. Section 6 deals with minimization of waiting costs w.r.t the mean values of the time limits, possibly with constraints on individual time limits and on the sum of all time limits. Some of the results are compared with the optimal values of job limits for similar systems. The main conclusions are summarized in Section 7.

2 Description of the model

The polling system consists of S stations and a single server. Jobs arrive at station j according to a Poisson process with rate λ_j, $j = 1, \ldots, S$. The superposition of the arrival processes at the various stations is a Poisson process with rate $\Lambda \doteq \sum_{j=1}^{S} \lambda_j$. Each queue may contain an unbounded number of jobs. At each station jobs are served in order of arrival. Service times of jobs arriving at station j have mean β_j, $j = 1, \ldots, S$. The load ρ_j offered at station j, $j = 1, \ldots, S$, and the total offered load ρ to the system are defined by

$$\rho_j \doteq \lambda_j \beta_j, \qquad \rho \doteq \sum_{j=1}^{S} \rho_j. \tag{2.1}$$

The server visits the stations in a fixed cyclic order $1, 2, \ldots, S, 1, \ldots$. The number of services which may be performed during a visit of the server to station j is determined by a time limit τ_j, $j = 1, \ldots, S$. As long as this time limit has not expired the server is allowed to start new services during a visit. A visit to a station ends either when a service is completed and the time limit has been exceeded, or when the station is or becomes empty. In practice, the time limits will be constant. However, in order to construct a Markov process with discrete state space these time limits will be approximated by random variables with Erlang distributions. The number of phases of the Erlang distribution of the time limit for a visit of the server to station j will be denoted by Γ_j, and the transition rates at the phases are $\gamma_j \doteq \Gamma_j/\tau_j$, so that the mean value of the Erlang distributed time limit is τ_j, $j = 1, \ldots, S$.

The times the server needs for switching from station $j-1$ to station j have means δ_j, $j =$

$1, \ldots, S$. The total mean switching time of the server during a tour along the stations will be denoted by $\Delta \doteq \sum_{j=1}^{S} \delta_j$.

From the general result on stability of polling systems in Fricker and Jaïbi [9] it follows that the present system is stable iff

$$\rho + \Delta \max_{j=1,\ldots,S} \{\lambda_j / G_j\} < 1; \tag{2.2}$$

here, G_j denotes the mean of the maximal number of jobs that can be served at station j during a visit of the server, $j = 1, \ldots, S$. The expression for this quantity for a station with a time limit is not as simple as that for a station with a job limit or a Bernoulli schedule. Let $T_j(t)$ denote the distribution of the random variable that determines when the time limit at station j expires, and let $B_j(t)$ denote the service time distribution at station j, $j = 1, \ldots, S$. Then it follows by conditioning on the realisations of the timer and of the service times that

$$G_j = \int_0^\infty \sum_{m=1}^{\infty} m \left[B_j^{(m-1)*}(t) - B_j^{m*}(t) \right] dT_j(t), \quad j = 1, \ldots, S. \tag{2.3}$$

3 The balance equations

It will be assumed throughout this paper that the polling systems are in steady state. The random variable N_j will indicate the number of jobs present at station j, $j = 1, \ldots, S$. Beside the vector of random variables $\mathbf{N} \doteq (N_1, \ldots, N_S)$ several supplementary variables are needed to obtain a Markov process. The supplementary variable H will indicate the station to which the server is switching or to which the server is attending. The supplementary variable Z will indicate the status of the server. More precisely, $Z = 0$ will indicate that the server is switching and $Z = \psi$ will indicate that the server is serving jobs while the timer is in phase ψ, $\psi = 1, \ldots, \Gamma_H + 1$; here, $Z = \Gamma_H + 1$ indicates that the timer has already expired during the current visit. For conciseness and clarity, the PSA will be described for the case of exponential service times, with rate $\mu_j \doteq 1/\beta_j$ at station j, and exponential switching times with rates $\nu_j \doteq 1/\delta_j$, $j = 1, \ldots, S$. Therefore, we do not need a variable to indicate the current phase of these distributions. The reader is referred to [4] for a description of the PSA for this model with Coxian distributions and general periodic polling orders. For the same reason, the PSA is presented without the use of a conformal mapping which is often necessary to avoid numerical inaccuracies, cf. [2]. In order to formulate the balance equations for the Markov process $(\mathbf{N}, H, Z)$ we will use the indicator function $I_{\{C\}}$ taking the values 0 (if C is false) or 1 (if C is

true), and the unit vectors $\mathbf{e}_j$, $j = 1, \ldots, S$, in $\mathbb{N}^S$.

The balance equations for the probabilities of states in which the server is switching are, for $\mathbf{n} \in \mathbb{N}^S$, $j = 1, \ldots, S$,

$$[\Lambda + \nu_j]p(\mathbf{n}, j, 0) = \sum_{h=1}^{S} \lambda_h I_{\{n_h \geq 1\}} p(\mathbf{n} - \mathbf{e}_h, h, 0) + \nu_{j-1} I_{\{n_{j-1}=0\}} p(\mathbf{n}, j-1, 0)$$
$$+ \mu_{j-1} p(\mathbf{n} + \mathbf{e}_{j-1}, j-1, \Gamma_{j-1}+1) + \mu_{j-1} I_{\{n_{j-1}=0\}} \sum_{\psi=1}^{\Gamma_{j-1}} p(\mathbf{n} + \mathbf{e}_{j-1}, j-1, \psi). \quad (3.1)$$

The first term at the righthand side stands for transitions caused by an arrival of a job at one of the stations. The second term describes a transition from a switch to station $j-1$ to a switch to station j; such a transition can only occur if station $j-1$ is empty. The third and fourth term describe a transition from a last service at station $j-1$ to a switch to station j; in the third term, the end of the visit is due to the expiration of the timer at station $j-1$, in the fourth one to the exemption of station $j-1$.

The balance equations for the probabilities of states in which the server is serving jobs are, for $\mathbf{n} \in \mathbb{N}^S$, $j = 1, \ldots, S$, $n_j \geq 1$, $\psi = 1, \ldots, \Gamma_j + 1$,

$$[\Lambda + \mu_j + \gamma_j I_{\{\psi \leq \Gamma_j\}}]p(\mathbf{n}, j, \psi) = \sum_{h=1}^{S} \lambda_h I_{\{n_h \geq 1\}} p(\mathbf{n} - \mathbf{e}_h, h, \psi)$$
$$+ \nu_j I_{\{\psi=1\}} p(\mathbf{n}, j, 0) + \gamma_j I_{\{\psi \geq 2\}} p(\mathbf{n}, j, \psi - 1) + \mu_j I_{\{\psi \leq \Gamma_j\}} p(\mathbf{n} + \mathbf{e}_j, j, \psi). \quad (3.2)$$

The first term at the righthand side stands for transitions caused by an arrival of a job at one of the stations. The second term describes a transition from a switch to station j to the first service at station j (the timer starts in phase $\psi = 1$). The third term describes a phase transition of the timer. The fourth term describes a transition from one service at station j to another service at station j; such a transition can only occur if the timer had not expired before the new service started.

Finally, the sum of the probabilities over all states is equal to one:

$$\sum_{n_1=0}^{\infty} \cdots \sum_{n_S=0}^{\infty} \sum_{j=1}^{S} \sum_{\psi=0}^{\Gamma_j+1} p(\mathbf{n}, j, \psi) = 1. \quad (3.3)$$

4 The power-series algorithm

First, we write $\lambda_j = a_j \rho$, $j = 1, \ldots, S$, and $\Lambda = A\rho$ to obtain a parametrization of the model as a function of the total offered load ρ. Then, we introduce power-series expansions of the state

probabilities as functions of ρ: for all $\mathbf{n} \in \mathbb{N}^S$, $j = 1, \ldots, S$, $\psi = 0, 1, \ldots, \Gamma_j + 1$,

$$p(\mathbf{n}, j, \psi) = \rho^{|\mathbf{n}|} \sum_{k=0}^{\infty} \rho^k b(k; \mathbf{n}, j, \psi). \tag{4.1}$$

Here and below, we use the notation $|\mathbf{n}| \doteq n_1 + \cdots + n_S$. See [2, 3] for a motivation of (4.1). The expansions (4.1) are substituted into the Eqs. (3.1) and (3.2) for the state probabilities. Equating coefficients of corresponding powers of ρ on both sides of these equations leads to relations for the coefficients of the power-series expansions of the state probabilities.

The recurrence relations for the coefficients of the probabilities of states in which the server is switching are, for $k = 0, 1, 2, \ldots$, $\mathbf{n} \in \mathbb{N}^S$, $j = 1, \ldots, S$,

$$\begin{aligned}
\nu_j b(k; \mathbf{n}, j, 0) = {} & \sum_{h=1}^{S} a_h I_{\{n_h \geq 1\}} b(k; \mathbf{n} - \mathbf{e}_h, h, 0) - A I_{\{k \geq 1\}} b(k-1; \mathbf{n}, j, 0) \\
& + \nu_{j-1} I_{\{n_{j-1}=0\}} b(k; \mathbf{n}, j-1, 0) + \mu_{j-1} I_{\{k \geq 1\}} b(k-1; \mathbf{n} + \mathbf{e}_{j-1}, j-1, \Gamma_{j-1}+1) \\
& + \mu_{j-1} I_{\{n_{j-1}=0\}} I_{\{k \geq 1\}} \sum_{\psi=1}^{\Gamma_{j-1}} b(k-1; \mathbf{n} + \mathbf{e}_{j-1}, j-1, \psi).
\end{aligned} \tag{4.2}$$

The recurrence relations for the coefficients of the probabilities of states in which the server is serving jobs are, for $k = 0, 1, 2, \ldots$, $\mathbf{n} \in \mathbb{N}^S$, $j = 1, \ldots, S$, $n_j \geq 1$, $\psi = 1, \ldots, \Gamma_j + 1$,

$$\begin{aligned}
[\mu_j + \gamma_j I_{\{\psi \leq \Gamma_j\}}] b(k; \mathbf{n}, j, \psi) = {} & \sum_{h=1}^{S} a_h I_{\{n_h \geq 1\}} b(k; \mathbf{n} - \mathbf{e}_h, h, \psi) - A I_{\{k \geq 1\}} b(k-1; \mathbf{n}, j, \psi) \\
& + \nu_j I_{\{\psi=1\}} b(k; \mathbf{n}, j, 0) + \gamma_j I_{\{\psi \geq 2\}} b(k; \mathbf{n}, j, \psi - 1) + \mu_j I_{\{\psi \leq \Gamma_j, k \geq 1\}} b(k-1; \mathbf{n} + \mathbf{e}_j, j, \psi).
\end{aligned} \tag{4.3}$$

Although the detailed form of Eqs. (4.2) and (4.3) is quite different from the equations for polling systems with job limits or Bernoulli schedules, their main structure in terms of k and $\mathbf{n}$ is the same. Therefore, Eqs. (4.2) and (4.3) can be used to compute the coefficients of the power-series expansions of the state probabilities recursively, cf. [2, 3], using the same order of computation as for other polling systems. The only case in which the coefficients can not be computed recursively is the case $\mathbf{n} = \mathbf{0}$; this is the only situation in which the server can make a complete tour along the stations without any change in the values of $\mathbf{N}$. In the case $\mathbf{n} = \mathbf{0}$ Eq. (4.2) reduces to a dependent set of equations for the coefficients $b(k; \mathbf{0}, j, 0)$, $j = 1, \ldots, S$, for each fixed k, $k = 0, 1, \ldots$. These small sets of equations can be solved together with the following relation which follows from Eq. (3.3): for $k = 0, 1, \ldots$,

$$\sum_{j=1}^{S} b(k; \mathbf{0}, j, 0) = I_{\{k=0\}} - \sum_{\substack{n_1=0 \\ 1 \leq n_1 + \cdots + n_S \leq k}}^{k} \cdots \sum_{n_S=0}^{k} \sum_{j=1}^{S} \sum_{\psi=0}^{\Gamma_j+1} b(k - |\mathbf{n}|; \mathbf{n}, j, \psi). \tag{4.4}$$

Table 1: Three-station model with varying number of phases of the timers.

Γ_1	Γ_2	Γ_3	$\mathcal{V}$	$E\{W_1\}$	$E\{W_2\}$	$E\{W_3\}$	$E\{W\}$	$\sigma\{W_1\}$	$\sigma\{W_2\}$	$\sigma\{W_3\}$	$\sigma\{W\}$
1	1	1	18	2.200	5.724	5.802	3.625	2.612	8.664	8.749	6.121
2	2	2	21	2.060	5.855	5.955	3.598	2.370	8.804	8.905	6.187
4	4	4	27	1.978	5.937	6.056	3.585	2.213	8.893	9.010	6.234
8	8	8	39	1.934	5.981	6.113	3.579	2.121	8.945	9.073	6.263
80	80	80	255	1.895	6.020	6.165	3.574	2.032	8.998	9.138	6.293
2	1	1	19	2.056	5.903	5.984	3.611	2.385	8.873	8.959	6.232
4	2	2	23	1.979	5.960	6.065	3.592	2.228	8.930	9.036	6.258
8	4	4	31	1.936	5.992	6.116	3.583	2.132	8.964	9.085	6.275
16	8	8	47	1.914	6.008	6.142	3.578	2.079	8.983	9.113	6.285
120	60	60	255	1.894	6.023	6.167	3.574	2.029	9.003	9.143	6.296
4	1	1	21	1.978	6.003	6.088	3.605	2.251	8.994	9.084	6.300
8	2	2	27	1.938	6.013	6.122	3.590	2.150	8.999	9.109	6.298
16	4	4	39	1.916	6.019	6.144	3.582	2.090	9.001	9.125	6.296
32	8	8	63	1.904	6.022	6.156	3.578	2.057	9.002	9.134	6.296
160	40	40	255	1.894	6.023	6.168	3.574	2.029	9.003	9.143	6.296
8	1	1	25	1.939	6.054	6.142	3.603	2.177	9.060	9.153	6.338
16	2	2	35	1.919	6.039	6.150	3.589	2.110	9.035	9.148	6.318
32	4	4	55	1.906	6.032	6.158	3.582	2.069	9.020	9.146	6.307
64	8	8	95	1.899	6.028	6.163	3.578	2.047	9.012	9.145	6.301
192	24	24	255	1.894	6.025	6.167	3.575	2.031	9.007	9.145	6.298

The coefficients of the power-series expansions of the moments are obtained from those of the state probabilities in a straightforward manner, cf. [3]. The moments of the waiting time distributions for jobs at the various stations, assuming service in order of arrival, follow from the moments of the marginal queue-length distribution through a general relationship between the generating function of the queue-length distribution and the Laplace-Stieltjes transform of the waiting time in M/G/1-type systems, cf. [12], formula (4.31). The waiting time of a job at station j is denoted by W_j, $j = 1, \ldots, S$, and W stands for the waiting time of an arbitrary job. The standard deviation of a random variable X will be denoted by $\sigma\{X\}$.

In Table 1 the influence of the number of phases of the Erlang distributions of the timers is illustrated for an example taken from [10]. The model consists of three stations. The arrival

Table 2: Estimation of the performance of the system with constant timers.

$\Gamma_1^{(a)}$	$\Gamma_2^{(a)}$	$\Gamma_3^{(a)}$	$\Gamma_1^{(b)}$	$\Gamma_2^{(b)}$	$\Gamma_3^{(b)}$	$E\{W_1^{(\mathrm{Det})}\}$	$E\{W_2^{(\mathrm{Det})}\}$	$E\{W_3^{(\mathrm{Det})}\}$	$E\{W^{(\mathrm{Det})}\}$
1	1	1	2	2	2	1.920	5.986	6.108	3.571
2	2	2	4	4	4	1.896	6.019	6.157	3.572
2	1	1	4	2	2	1.902	6.017	6.146	3.573
4	2	2	8	4	4	1.893	6.024	6.167	3.574
4	1	1	8	2	2	1.898	6.023	6.156	3.575
8	2	2	16	4	4	1.894	6.025	6.166	3.576

rates are $\lambda_1 = 0.6$, $\lambda_2 = \lambda_3 = 0.2$, the service times are exponential with means $\beta_j = 0.8$, $j = 1, 2, 3$, and the switching times are Erlang E_4 distributed with means $\delta_j = 0.05$, $j = 1, 2, 3$. The quantity $\mathcal{V}$ denotes the size of the supplementary space, in this case $3 \times 4 + 3 \sum_{j=1}^{3} (\Gamma_j + 1)$. The mean values of the timers are $\tau_1 = 23.2$, $\tau_2 = \tau_3 = 1.6$, so that on the average 30 jobs can be served during a visit of the server to station 1 and 3 jobs during visits to both station 2 and 3. We have selected this model, because [10] reports the largest differences in mean waiting times between systems with exponential timers and constant timers in this example. In [10] this model has been solved with constant switching times and exponential timers; the reported mean waiting times are $E\{W_1\} = 2.198$, $E\{W_2\} = 5.713$ and $E\{W_3\} = 5.792$. Our results with E_4 distributed switching times and exponential timers are close to these values. [10] also reports simulation results for this model with constant switching times and constant timers: $E\{W_1\} = 1.884 \pm 0.020$, $E\{W_2\} = 5.989 \pm 0.132$ and $E\{W_3\} = 6.163 \pm 0.148$. Our results with E_4 distributed switching times and Erlang distributed timers show that a rather large number of phases of the corresponding Erlang distribution is required to obtain a close approximation for the performance measures of a similar system but with constant timers, when the time limit at a station is relatively large compared to the mean service time. However, it is shown in Table 2 that simple linear extrapolations in the squared coefficients of variation of the Erlang distributions of the timers yield good approximations for the performance measures of the system with constant timers, based on the evaluation of two systems with Erlang distributed timers with only a few phases. In the table, each row shows two sets of numbers of phases of the Erlang distributions of the timers, indicated by (a) and (b). For instance, the means of performance measures of the system with constant timers are estimated from those with the

Erlang timers by the simple extrapolation

$$E\{X^{(\mathrm{Det})}\} \approx E\{X^{(b)}\} - [E\{X^{(a)}\} - E\{X^{(b)}\}] = 2E\{X^{(b)}\} - E\{X^{(a)}\}. \tag{4.5}$$

Here, X is some performance measure, $X^{(\mathrm{Det})}$ indicates the version with deterministic timers, and $X^{(a)}$ and $X^{(b)}$ stand for the versions with Erlang timers. A similar extrapolation can be applied to standard deviations, cf. Table 1.

5 Derivatives with the PSA

For optimization of a performance measure with respect to real-valued parameters of a system it is useful to be able to compute derivatives of the performance measure as function of these parameters. Then, optimization techniques as the conjugate gradient method can be used to determine optimal values of these parameters with respect to some objective function. For the present model, one might be interested in determining optimal time limits w.r.t. some objective. Because the (mean) time limits do not occur explicitly in Eqs. (4.2) and (4.3), we consider derivatives w.r.t. the transition rates of the timers γ_j, $j = 1, \ldots, S$. It can be shown that these derivatives possess power-series expansions of the form, cf. Eq. (4.1): for all $\mathbf{n} \in \mathbb{N}^S$, $j, r = 1, \ldots, S$, $\psi = 0, 1, \ldots, \Gamma_j + 1$,

$$\frac{\partial}{\partial \gamma_r} p(\mathbf{n}, j, \psi) = \rho^{|\mathbf{n}|} \sum_{k=0}^{\infty} \rho^k b_r(k; \mathbf{n}, j, \psi). \tag{5.1}$$

Taking derivatives in Eqs. (4.2) and (4.3) with respect to γ_j, $j = 1, \ldots, S$, leads to a further set of recursions. The recurrence relations for the coefficients of the derivatives of the probabilities of states in which the server is switching are exactly the same as those for the corresponding probabilities, because the rates γ_j, $j = 1, \ldots, S$, do not occur in Eq. (4.2). Such a property does not hold for polling systems with Bernoulli schedules, cf. [3]. The recurrence relations for the coefficients of the derivatives of the probabilities of states in which the server is serving jobs are, for $k = 0, 1, 2, \ldots$, $\mathbf{n} \in \mathbb{N}^S$, $j, r = 1, \ldots, S$, $n_j \geq 1$, $\psi = 1, \ldots, \Gamma_j + 1$,

$$\begin{aligned}
[\mu_j + \gamma_j I_{\{\psi \leq \Gamma_j\}}] b_r(k; \mathbf{n}, j, \psi) &= \sum_{h=1}^{S} a_h I_{\{n_h \geq 1\}} b_r(k; \mathbf{n} - \mathbf{e}_h, h, \psi) - A I_{\{k \geq 1\}} b_r(k-1; \mathbf{n}, j, \psi) \\
&+ \nu_j I_{\{\psi = 1\}} b_r(k; \mathbf{n}, j, 0) + I_{\{r=j\}} [I_{\{\psi \geq 2\}} b(k; \mathbf{n}, j, \psi - 1) - I_{\{\psi \leq \Gamma_j\}} b(k; \mathbf{n}, j, \psi)] \\
&+ \gamma_j I_{\{\psi \geq 2\}} b_r(k; \mathbf{n}, j, \psi - 1) + \mu_j I_{\{\psi \leq \Gamma_j\}} I_{\{k \geq 1\}} b_r(k-1; \mathbf{n} + \mathbf{e}_j, j, \psi).
\end{aligned} \tag{5.2}$$

The law of total probability leads to a similar relation as Eq. (4.4) for the derivatives. Note that the coefficients $b_r(0;\mathbf{0},j,\psi)$ do not vanish for $\psi \geq 1$, in contrast with those of polling models with Bernoulli schedules, cf. [3].

The coefficients $b_r(k;\mathbf{n},j,\psi)$ can be computed recursively, but only in conjunction with the coefficients $b(k;\mathbf{n},j,\psi)$. Derivatives of performance measures with respect to the (mean) time limits can be computed from the above: for all $\mathbf{n} \in \mathbb{N}^S$, $j,r = 1,\ldots,S$, $\psi = 0,1,\ldots,\Gamma_j+1$,

$$\frac{\partial}{\partial \tau_r} p(\mathbf{n},j,\psi) = -\frac{\Gamma_r}{\tau_r^2} \frac{\partial}{\partial \gamma_r} p(\mathbf{n},j,\psi). \tag{5.3}$$

6 Optimization of the time limits

Consider the following optimization problem with the time limits as decision variables:

$$\min_{\tau_1,\ldots,\tau_S} C \doteq \sum_{j=1}^{S} c_j \lambda_j E\{W_j\}, \tag{6.1}$$

subject to

$$L_j \leq \tau_j \leq U_j, \quad j = 1,\ldots,S; \tag{6.2}$$

$$\sum_{j=1}^{S} \tau_j \leq B. \tag{6.3}$$

The coefficients c_j in the objective function indicate the relative waiting cost for jobs at station j, $j = 1,\ldots,S$. Note that $\tau_j \downarrow 0$ implies $\gamma_j \to \infty$, $j = 1,\ldots,S$. If a time limit vanishes no job would ever be served at the corresponding station. We have taken $L_j = 10^{-6}$, $j = 1,\ldots,S$, in all examples to prevent that γ_j becomes too large. A very small but positive value of a time limit means in fact that the server is allowed to serve exactly one job during each visit to the corresponding station. On the other hand, a large value of a time limit means that the corresponding station is served exhaustively. In the cases that we indicate that the optimal time limit is infinite the optimization procedure stopped at some finite value of that time limit because the derivative of the cost function with respect to that time limit became too small. We have also evaluated the cost function in those cases with a much larger value of the time limit, and have found no significant differences in costs.

Table 3 shows the unconstrained optimal time limits as function of the load for a three-station system with the following parameters: arrival rates $\lambda_1 = \frac{2}{3}\rho$, $\lambda_2 = \lambda_3 = \frac{1}{3}\rho$, Erlang E_2 distributed service times with means $\beta_1 = 1.0$, $\beta_2 = \beta_3 = 0.5$, Erlang E_2 distributed switching times with means $\delta_j = 0.1$, $j = 1,2,3$, and cost factors $c_1 = 1.2$, $c_2 = c_3 = 0.3$. The optimal

Table 3: Optimal time limits as function of the load.

	Exponential timers					Erlang-2 timers					Erlang-4 timers			
ρ	τ_1^*	τ_2^*	τ_3^*	C_{M}	C_{E_4}	τ_1^*	τ_2^*	τ_3^*	C_{E_2}	C_{E_4}	τ_1^*	τ_2^*	τ_3^*	C_{E_4}
0.3	∞	∞	∞	0.139	0.139	∞	∞	∞	0.139	0.139	∞	∞	∞	0.139
0.4	∞	∞	∞	0.246	0.246	∞	2.98	4.14	0.246	0.246	∞	1.78	2.09	0.246
0.5	∞	2.95	4.14	0.414	0.413	∞	1.68	1.96	0.413	0.413	∞	1.34	1.53	0.413
0.6	∞	1.99	2.19	0.681	0.679	∞	1.46	1.59	0.679	0.679	∞	1.26	1.35	0.678
0.7	∞	2.06	2.02	1.147	1.140	∞	1.60	1.64	1.141	1.138	∞	1.42	1.47	1.138
0.8	∞	2.82	2.49	2.096	2.078	∞	2.17	2.14	2.081	2.074	∞	1.89	1.95	2.073
0.9	∞	5.82	4.36	4.955	4.897	∞	4.16	3.83	4.909	4.881	∞	3.84	3.83	4.881

time limits are shown for the case of exponentially distributed timers as well as for the case of Erlang E_2 respectively Erlang E_4 distributed timers at all stations. In the cases of exponentially and Erlang E_2 distributed timers (with minimal cost indicated by C_{M} respectively C_{E_2}) the system has also been evaluated with the same values of the mean time limits, but with Erlang E_4 distributed timers (cost C_{E_4}). The results in the table illustrate that the cost functions are rather flat near their minima, and that optimization is less sensitive to the number of phases of the timers than evaluation of performance measures. Note the resemblance of the behaviour of the optimal time limits as function of the load with that of the optimal Bernoulli schedules in [5]. In particular, the optimal set of time limits is, for each set of cost factors, such that at least one time-limit is infinite (i.e., the corresponding station is served exhaustively), and the stations for which the time limits are infinite are the stations for which the ratio $c_j\mu_j$ is maximal over $j = 1, \ldots, S$ (in agreement with the '$c\mu$'-rule for priority systems). For the other stations it holds that the optimal time limit tends to infinity in light traffic ($\rho \downarrow 0$) as well as in heavy traffic ($\rho \uparrow 1$). In light traffic, finite time limits might force the server to make an often unnecessary tour along the stations to search for jobs which will only be present with small probability. In heavy traffic, the time limits have to be large in order to keep the system stable, i.e., to compensate for the loss of server availability due to the switching times.

Our final example concerns the cyclic polling system with five stations which has been considered in [7], Tables III and IX. The arrival rates are $\lambda_1 = 0.35$, $\lambda_2 = \ldots = \lambda_5 = 0.10$, the service times are exponential with means $\beta_j = 1.0$, $j = 1, \ldots, 5$, and the switching times are exponential with means $\delta_2 = 0.10$, $\delta_j = 0.05$, $j = 1, 3, 4, 5$. Hence, $\rho = 0.75$ and $\Delta = 0.30$. The cost factor for

Table 4: Optimal time limits and job limits for a five-station model without constraint.

c_{2-5}	τ_1^*	τ_2^*	τ_3^*	τ_4^*	τ_5^*	C	$C^{(\mathrm{Det})}$	K_1^*	K_2^*	K_3^*	K_4^*	K_5^*	C
0.1	∞	0.00	0.00	0.00	0.00	0.88	0.88	∞	1	1	1	1	0.88
0.5	∞	1.41	1.37	1.34	1.32	1.76	1.74	∞	2	2	2	2	1.76
1.0	∞	∞	∞	∞	∞	2.63	2.63	∞	∞	∞	∞	∞	2.63
2.0	2.83	∞	∞	∞	∞	3.94	3.86	3	∞	∞	∞	∞	3.90
10.0	0.34	∞	∞	∞	∞	10.47	10.40	1	∞	∞	∞	∞	10.74

station 1 is fixed, $c_1 = 1.0$. The cost factors for the other stations are equal, and are denoted by $c_{2-5} \doteq c_2 = \ldots = c_5$. Table 4 shows the optimal time limits and the minimal cost for the unconstrained optimization problem with exponentially distributed timers, for several values of c_{2-5}. For comparison, this table also contains the optimal job limits K_j^*, $j = 1, \ldots, 5$, and the corresponding minimal cost. A job limit places a maximum on the *number* of jobs which may be served during a visit of the server to a station. The optimal set of job limits can only be determined by enumeration of all, infinitely many, sets of job limits. In [7], Table IX, the supposedly optimal sets of job limits have been found by a limited enumeration and on the basis of a conjecture which implies that at least one of the optimal job limits is infinite, which means that the corresponding station is served exhaustively. This conjecture is supported by the numerical results on systems with Bernoulli schedules in [5]. In contrast with Bernoulli parameters (probabilities) with their finite range time limits have an infinite range. Therefore, numerical procedures for optimization of systems with unconstrained time limits will stop at some finite value for all time limits. In cases of very large values of a time limit we have compared the cost of the found set of finite time limits with the cost corresponding a similar set of time limits but where the stations with an originally large time limit are served exhaustively. The so obtained results also confirm the conjecture in [7]. In the example with $c_{2-5} = 2.0$ in Table 4 the minimal cost of 3.94 attainable with exponentially distributed timers is larger than the minimal cost of 3.90 attainable with job limits. However, when we apply this set of mean time limits with Erlang distributed timers then the cost reduces to 3.89 with an Erlang E_4 distributed timer and to an estimated 3.86 with a constant time limit, cf. Eq. (4.5). The estimated minimal costs for the case of constant timers are indicated in Table 4 in the column with the header $C^{(\mathrm{Det})}$. It seems that the larger flexibility in adjusting the time limits allows a

Table 5: Optimal time limits and job limits for a five-station model with a constraint.

c_{2-5}	τ_1^*	τ_2^*	τ_3^*	τ_4^*	τ_5^*	C	$C^{(\mathrm{Det})}$	K_1^*	K_2^*	K_3^*	K_4^*	K_5^*	C
0.1	15.00	0.00	0.00	0.00	0.00	1.00	0.90	16	1	1	1	1	0.88
0.5	13.95	0.30	0.27	0.25	0.23	1.87	1.79	12	2	2	2	2	1.77
1.0	8.14	1.73	1.72	1.71	1.70	2.81	2.75	8	3	3	3	3	2.67
2.0	1.94	3.28	3.27	3.26	3.25	4.15	3.99	3	4	4	4	5	3.91
10.0	0.22	3.71	3.70	3.69	3.68	11.18	10.78	1	4	5	5	5	10.75

lower minimal cost than that realizable with job limits which are restricted to integer values. Table 5 shows the optimal sets of job limits and the corresponding minimal costs for the same system, but with the constraint $\sum_{j=1}^{5} K_j \leq 20$ on the total number of services that the server is allowed to perform during a cycle. These results are based on $M = 29$ terms of the power-series expansions; the estimated errors are in the order of 1%, much more than the estimated errors in the performance measures for the systems with time limits. Our results deviate in some cases from those reported in [7], Table III. Table 6 contains more elaborate data on the costs as functions of the job limits. Note that $E\{W_1\}$ increases with decreasing K_1, while the other mean waiting times show a zigzag behavior for $K_1 \geq 8$, while they are decreasing with decreasing K_1 for $K_1 < 8$. This non-monotonic behavior is due to the integer nature of the job limits. The case $c_{2-5} = 10.0$ which we have added reveals that it may be far from optimal in the constraint case to restrict the search for the set of optimal job limits to those sets for which $K_2 = \ldots = K_5$, although all parameters related to stations 2–5 are equal. In this particular case it would lead to a cost of 13.27, 23% more than the minimal cost of 10.74. Because the service times are exponentially distributed in this model, the mean of the maximal number of services per visit to station j is $1 + \tau_j/\beta_j$, cf. Eq. (2.3), when a mean time limit τ_j is applied, $j = 1, \ldots, 5$. Therefore, we have determined optimal mean time limits for the case of exponential timers under the constraint $\sum_{j=1}^{5} \tau_j \leq 15$, for comparison with optimal sets of job limits. The results are displayed in Table 5. We have also determined the optimal time limits for the case of Erlang E_2 distributed timers. The corresponding minimal costs are somewhat less than the minimal costs corresponding to the case of exponential timers. The estimated minimal costs for the case of constant timers, $C^{(\mathrm{Det})}$, are still somewhat higher than the corresponding minimal costs with job limits. This feature must be due to the randomness

Table 6: Performance of a five-station model with job limits.

job limits					mean waiting times					waiting cost C with c_{2-5} =				
K_1	K_2	K_3	K_4	K_5	W_1	W_2	W_3	W_4	W_5	0.1	0.5	1.0	2.0	10.0
16	1	1	1	1	1.87	5.72	5.73	5.75	5.78	0.88	1.80	2.95	5.25	23.62
15	1	1	1	2	1.96	6.20	6.22	6.24	3.52	0.91	1.80	2.90	5.12	22.86
14	1	1	2	2	2.07	6.81	6.82	3.80	3.84	0.94	1.79	2.85	4.98	21.99
13	1	2	2	2	2.20	7.62	4.17	4.20	4.24	0.97	1.78	2.79	4.82	21.00
12	2	2	2	2	2.37	4.66	4.70	4.73	4.77	1.02	1.77	2.71	4.60	19.69
11	2	2	2	3	2.49	4.81	4.84	4.87	3.80	1.05	1.79	2.70	4.54	19.20
10	2	2	3	3	2.61	4.95	4.98	3.85	3.88	1.09	1.80	2.68	4.45	18.58
9	2	3	3	3	2.85	5.09	3.89	3.91	3.96	1.17	1.84	2.68	4.37	17.85
8	3	3	3	3	3.10	3.91	3.94	3.98	4.00	1.24	1.88	2.67	4.25	16.92
7	3	3	3	4	3.38	3.80	3.83	3.86	3.47	1.33	1.93	2.68	4.17	16.14
6	3	3	4	4	3.73	3.62	3.65	3.34	3.37	1.45	2.00	2.70	4.10	15.28
5	3	4	4	4	4.15	3.39	3.13	3.16	3.18	1.58	2.10	2.74	4.03	14.32
4	4	4	4	4	4.58	2.89	2.90	2.92	2.95	1.72	2.19	2.77	3.94	13.27
3	4	4	4	5	5.23	2.59	2.60	2.63	2.60	1.93	2.35	2.87	3.91	12.25
2	4	4	5	5	6.35	2.22	2.24	2.23	2.27	2.31	2.67	3.12	4.01	11.17
1	4	5	5	5	10.12	1.77	1.77	1.81	1.86	3.61	3.90	4.26	4.98	10.75

of the service times. Finally, note that the optimal service disciplines in the case $c_{2-5} = 0.1$ are extremal. In such a case it might be profitable to consider generalized service disciplines in which, e.g., some stations are only served every second cycle.

Numerical experiments indicate that the optimization problems considered in this section possess a unique solution. In some cases, the objective function is very flat near its minimum. Both properties have also been found in [5] for the unconstrained optimization problem with Bernoulli schedules as service disciplines.

When using the PSA for optimization purposes it is often a good strategy for reducing computation time to start the search with a moderate number of terms of the power-series expansions, and then to improve the approximated optimum by using more terms. See [4] for further examples including one with a non-cyclic polling order.

7 Conclusions

Numerical results with the PSA show that many phases of Erlang distributed timers may be needed for accurate evaluation of systems with constant timers. However, good approximations for systems with constant timers can be obtained by extrapolation from results for systems with Erlang distributed timers with only a few phases. When minimizing the waiting cost by optimizing the values of the timers the use of a small number of phases often yields timer values with waiting cost close to minimal when used in systems with constant timers. The latter is partly due to the flatness of the cost function near its minimum.

References

[1] Baker, J.E. & Rubin, I. (1987). Polling with a general-service order table. *IEEE Transactions on Communications* **35**: 283-288.

[2] Blanc, J.P.C. (1992). Performance evaluation of polling systems by means of the power-series algorithm. *Annals of Operations Research* **35**: 155-186.

[3] Blanc, J.P.C. (1993). Performance analysis and optimization with the power-series algorithm. In L. Donatiello & R. Nelson (eds.), *Performance Evaluation of Computer and Communication Systems*. Berlin: Springer, pp. 53-80.

[4] Blanc, J.P.C. Optimization of periodic polling systems with non-preemptive, time-limited service, CentER Discussion paper 9663, Tilburg University, 1996.

[5] Blanc, J.P.C. & Van der Mei, R.D. (1995). Optimization of polling systems with Bernoulli schedules, *Performance Evaluation* **22**: 139-158.

[6] Blanc, J.P.C. & Van der Mei, R.D. (1996). Computation of derivatives by means of the power-series algorithm, *INFORMS Journal on Computing* **8**: 45-54.

[7] Borst, S.C., Boxma, O.J., & Levy, H. (1995). The use of service limits for efficient operation of multistation single-medium communication systems, *IEEE/ACM Transactions on Networking* **3**: 602-612.

[8] Coffman, E. Jr., Mitrani, I., & Fayolle, G. (1988). Two queues with alternating service periods. In. P.-J. Courtois, G. Latouche (eds.), *Performance '87*. Amsterdam: North-Holland, pp. 227-239.

[9] Fricker, C., M.R. Jaïbi. (1994). Monotonicity and stability for periodic polling models, *Queueing Systems* **15**: 211-238.

[10] Leung, K.K. (1994). Cyclic-service systems with nonpreemptive, time-limited service. *IEEE Transactions on Communications* **42**: 2521-2524.

[11] Resing, J.A.C. Polling systems and multitype branching processes, *Queueing Systems* **13** (1993), 409–426.

[12] Takagi, H. *Analysis of Polling Systems.* The MIT Press, 1986.

[13] Van den Hout, W.B. & Blanc, J.P.C. (1995) The power-series algorithm for Markovian queueing networks. In W.J. Stewart (ed.), *Computations with Markov Chains.* Boston: Kluwer, pp. 321-338.